FastR: Fast database search tool for structured RNA sequences[1]

Vineet Bafna[2], Shaojie Zhang[2]

Keywords: non-coding genes, RNA, database search, filtration, dynamic programming

1 Introduction.

The discovery of novel non-coding RNAs has been among the most exciting recent developments in Biology. Yet, many more remain undiscovered. It has been hypothesized that there is in fact an *abundance* of functional non-coding RNA (ncRNA) with various catalytic and regulatory functions [1]. Computational methods tailored specifically for ncRNA are being actively developed. As the inherent signal for ncRNA is weaker than that for protein coding genes, comparative methods offer the most promising approach, and are the subject of our research.

We consider the following problem: Given an RNA sequence with a known secondary structure, efficiently compute all structural homologs (computed as a function of sequence and structural similarity) in a genomic database. Our approach, based on structural filters that eliminate a large portion of the database, while retaining the true homologs allows us to search a typical bacterial database in minutes on a standard PC, with high sensitivity and specificity. This is two orders of magnitude better than current available software for the problem.

2 Methods and Results

Recently, Klein and Eddy [3] developed a tool, RSEARCH, for searching a database with a query RNA molecule. The method depends upon existing algorithms for computing alignments between an RNA sequence and substrings of a database, where the alignment score is a function of sequence and structural similarity. Known algorithms for computing such alignments are computationally intensive (approximately $O(mw^2n)$, where m is the length of the query sequence, n is the length of the database sequence, and w is the maximum length of a database substring that is aligned to the query). Not surprisingly, RSEARCH is slow to use. For a test run on an Intel/linux PC with 2.8GHz, 1Gb memory, a microbial database of size 1.67M, and a query 5S rRNA sequence, the program took over 6.5 hrs. to run. This makes it impractical when either the query or the database is large.

We propose *FastR* , an efficient database search tool for ncRNA. An analogy can be drawn from fast search tools (BLAST/FASTA) for DNA and Protein sequences that has made database searching practical. The speed and effectiveness of BLAST in particular has contributed in large measure to the exponential growth of sequence databases, and the use of database search as an accepted method for finding novel DNA/protein homologs. By proposing FastR, which includes a novel idea for RNA structure filtering, and a novel & simple RNA alignment algorithm, we hope to do the same for ncRNA. As an example, FastR reduces the compute time of the previously mentioned query to 103s.

[1]The full-paper version of this work has been submitted to ISMB/ECCB 2004.

[2]Department of Computer Science and Engineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0114. Email: vbafna@cs.ucsd.edu; shzhang@cs.ucsd.edu

	Query	**Hits (TP/Tot)**	**Time**
RSEARCH	Asn-tRNA(AE001087.1/4936-5008)	85/93	3411s
FastR	”	71/87, 72/97	52s
RSEARCH	5S rRNA (AE016770.1/210436-210555)	97/97	14939s
FastR	”	79/99	44s
RSEARCH	Purine-Rs (AE010606.1/4680-4581)	33/39	9215s
FastR	”	27/35	30s
RSEARCH	Hammerhead (M83545.1/56-3)	50/58	2741s
FastR	”	44/51, 45/61	34s

Table 1: Comparison of FastR and RSEARCH.

Query	Genome	FastR (hits/TP/FN)	RSEARCH (E-val ≤ 10)	FastR time	RSEARCH time
Asn tRNA	*A. pernix*	25/24/9	57/31/2	2m57s	146m22s
5S rRNA	*A. pernix*	9/1/1	2/1/1	1m43s	390m7s

Table 2: Comparison of RSEARCH and FastR results on querying the 1.67Mb *A. pernix* genome (NC_000854.1). The true positives are obtained from known annotations. For False Negatives, we do not consider tRNAs with introns.

To test our algorithms, we worked with arbitrary ncRNA subfamilies of known/predicted structure from the RFAM [2] and the 5S Ribosomal RNA database [4]. Four sub-families are considered here, tRNA, 5S rRNA, a Purine Riboswitch, and the Hammerhead Ribozyme. For every sub-family, we chose some members arbitrarily, and inserted them in a random database of 1Mb, and tested our algorithms on the composite sequence. Table 1 summarizes the results of our search. As can be seen, FastR is close to two orders of magnitude faster than RSEARCH while maintaining comparable sensitivity.

We have also tested FastR on real genomes, where it is difficult to distinguish true hits. As shown in Table 2, querying the 1.67 Mb *A. pernix* genome yielded comparable results. FastR could not detect the 14 intron containing tRNAs, but detected 24 out of the remaining 33. For 5S rRNA, the single known annotation was the top hit, but there were other alignments of similar quality, indicative of novel 5S rRNAs. In the other two cases (Hammerhead and Purine-Rioboswitch), RSEARCH did not return any significant hit, and no annotations were available, hence no comparison could be made. FastR dominates again in speed.

References

[1] Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nature Reviews in Genetics*, 2: 919–929.

[2] Griffiths-Jones, S., Bateman, A., Marshall, M, Khanna, M. and S.R. Eddy. 2003. Rfam: an RNA family database. *NAR*, 31(1): 439–441.

[3] Klein, R.J. and Eddy, S.R. 2003. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics* 4(1): 44.

[4] Szymanski, M., Barciszewska, M.Z., Erdmann, V.A. and Barciszewski, J. 2000. 5S ribosomal RNA database. *NAR*, 28(1): 166–167.